Theoretical and practical considerations on the convergence properties of the Fisher-EM algorithm

Charles Bouveyron* & Camille Brunet†

* Laboratoire SAMM, EA 4543
Université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac, 75013 Paris, France

† Equipe Modal'X, EA 3454
Université Paris Ouest Nanterre La Defense
200 av. de la République, 92000 Nanterre, France

Abstract

The Fisher-EM algorithm has been recently proposed in [4] for the simultaneous visualization and clustering of high-dimensional data. It is based on a latent mixture model which fits the data into a latent discriminative subspace with a low intrinsic dimension. Although the Fisher-EM algorithm is based on the EM algorithm, it does not respect at a first glance all conditions of the EM convergence theory. Its convergence toward a maximum of the likelihood is therefore questionable. The aim of this work is two folds. Firstly, the convergence of the Fisher-EM algorithm is studied from the theoretical point of view. It is in particular proved that the algorithm converges under weak conditions in the general case. Secondly, the convergence of the Fisher-EM algorithm is considered from the practical point of view. It is shown that the Fisher's criterion can be used as stopping criterion for the algorithm to improve the clustering accuracy. It is also shown that the Fisher-EM algorithm converges faster than both the EM and CEM algorithm.

Keywords: high-dimensional data, model-based clustering, discriminative subspace, Fisher-EM algorithm, convergence properties.

1. Introduction

With the exponential growth of measurement capacities, the measured observations are nowadays frequently high-dimensional and clustering such

data remains a challenging problem. In particular, when considering the mixture model context, the corresponding clustering methods show a disappointing behavior in high-dimensional spaces. They suffer from the well-known *curse of dimensionality* [2] which is mainly due to the fact that model-based clustering methods are over-parametrized in high-dimensional spaces.

Fortunately, since the dimension of observed data is usually higher than their intrinsic dimension, it is theoretically possible to reduce the dimension of the original space without loosing any information. In the literature, a very common way to reduce the dimension is to use feature extraction methods such as principal component analysis (PCA) or feature selection methods. However, as shown by Chang [7], the principal components linked to the largest eigenvalues do not necessary contain the most relevant information about the group structure of the dataset. An alternative to dimension reduction methods is subspace clustering [5, 12, 13, 14, 16]. These techniques model the data of each group in low-dimensional subspaces while keeping all original dimension. Even though these methods turned out to be very efficient in practice, they are usually not able to provide a global visualization of the clustered data since they model each group in a specific subspace.

To overcome this limitation, Bouveyron and Brunet [4] recently proposed a new statistical framework which aims to simultaneously cluster the data and produce a low-dimensional representation of the clustered data. To that end, the proposed model clusters the data into a common latent subspace which both best discriminates the groups according to the current fuzzy partition of the data and has an intrinsic dimension lower than the dimension of the observation space. The proposed inference procedure for this latent mixture model is called the Fisher-EM algorithm. It is based on an EM procedure from which an additional step, named F-step, is introduced to estimate the projection matrix whose columns span the discriminative latent space. This projection matrix is estimated at each iteration by maximizing a constrained Fisher's criterion conditionally to the current soft partition of the data. As reported by [4], the Fisher-EM algorithm turns out to outperform most of the existing clustering and subspace clustering methods while providing in addition a useful visualization of the clustered data.

However, with the introduction of this additional step, the Fisher-EM algorithm does not satisfy at a first glance to all conditions required by the convergence theory of the EM algorithm. Indeed, the update of the orienta-

tion matrix in the F step is not done by directly maximizing the expected complete log-likelihood as required in the EM algorithm theory. From this point of view, the convergence toward a maximum of the likelihood of the Fisher-EM algorithm cannot be guaranteed and is therefore questionable.

This paper consequently focuses on the convergence properties of the Fisher-EM algorithm and is organized as follows. Section 2 reviews the discriminative latent mixture model and the Fisher-EM algorithm which was proposed for its inference. Section 3 focuses on theoretical aspects. The convergence of the Fisher-EM algorithm is in particular proved in two different cases. Numerical experiments are then presented in Section 4 to highlight the practical behavior of the convergence. Some concluding remarks and ideas for further works are finally given in Section 5.

2. The DLM model and the Fisher-EM algorithm

The discriminative latent mixture (DLM) model [4] aims to both cluster the data at hand and reduce their dimensionality into a common latent subspace. Conversely to similar approaches, such as [5, 13, 15, 16, 18], this latent subspace is assumed to be discriminative and its intrinsic dimension is strictly bounded by the number of groups.

2.1. The DLM model

Let $\{y_1, \ldots, y_n\} \in \mathbb{R}^p$ denote a dataset of n observations that one wants to cluster into K homogeneous groups, *i.e.* adjoin to each observation y_j a value $z_j \in \{1, \ldots, K\}$ where $z_i = k$ indicates that the observation y_i belongs to the kth group. On the one hand, let us assume that $\{y_1, \ldots, y_n\}$ are independent observed realizations of a random vector $Y \in \mathbb{R}^p$ and that $\{z_1, \ldots, z_n\}$ are also independent realizations of a random vector $Z \in \{1, \ldots, K\}$. On the other hand, let $\mathbb{E} \subset \mathbb{R}^p$ denote a latent space assumed to be the most discriminative subspace of dimension $d \leq K - 1$ such that $\mathbf{0} \in \mathbb{E}$ and where d is strictly lower than the dimension p of the observed space. Moreover, let $\{x_1, \ldots, x_n\} \in \mathbb{E}$ denote the actual data, described in the latent space $\mathbb{E}$ of dimension d, which are in addition presumed to be independent realizations of an unobserved random vector $X \in \mathbb{E}$. Finally, for each group, the observed variable $Y \in \mathbb{R}^p$ and the latent variable $X \in \mathbb{E}$ are assumed to be linked through a linear transformation:

$$Y = UX + \varepsilon, \tag{2.1}$$

where U is a $p \times d$ orthonormal matrix common to the K groups and satisfying $U^t U = \mathbf{I}_d$. The p-dimensional random vector ε stands for the noise term and, conditionally to Z, ε is assumed to be distributed according to a centered Gaussian density function with covariance matrix Ψ_k ($\varepsilon_{|Z=k} \sim \mathcal{N}(0, \Psi_k)$). Besides, within the latent space, X is assumed to be Gaussian conditionally to $Z = k$:

$$X_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{2.2}$$

where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are respectively the mean vector and the covariance matrix of the kth group. Given these distribution assumptions and according to equation (2.1),

$$Y_{|X,Z=k} \sim \mathcal{N}(UX, \Psi_k), \tag{2.3}$$

and its marginal distribution is therefore a mixture of Gaussians:

$$f(y) = \sum_{k=1}^{K} \pi_k \phi(y; m_k, S_k), \tag{2.4}$$

where π_k is the mixing proportion of the kth group and $\phi(.)$ denotes the multivariate Gaussian density function parametrized by the mean vector $m_k = U\mu_k$ and the covariance matrix $S_k = U\Sigma_k U^t + \Psi_k$ of the kth group. Furthermore, a $p \times p$ matrix $W = [U, V]$ is defined, satisfying the condition $W^t W = W W^t = \mathbf{I}_p$, where the $(p - d) \times p$ matrix V is an orthogonal complement of U. Finally, the noise covariance matrix Ψ_k is assumed to satisfy the conditions $V\Psi_k V^t = \beta_k \mathbf{I}_{p-d}$ and $U\Psi_k U^t = \mathbf{0}_d$, such that $\Delta_k = W^t S_k W$ has the following form:

$$\Delta_k = \begin{pmatrix} \boxed{\Sigma_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} \beta_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \beta_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \end{matrix}}\right\} d \leq K - 1 \\ \left.\vphantom{\begin{matrix} a \\ a \end{matrix}}\right\} (p - d) \end{matrix}$$

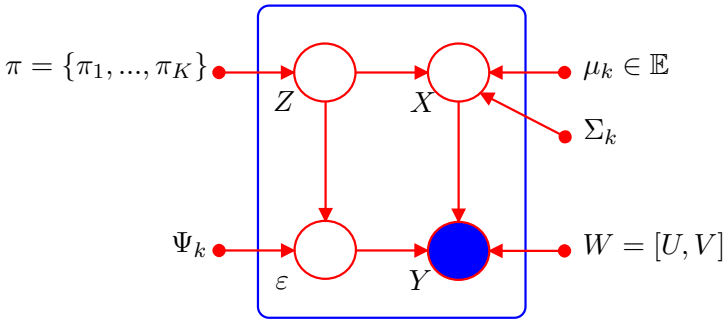

Figure 2.1: Graphical summary of the $\text{DLM}_{[\Sigma_k \beta_k]}$ model.

These last conditions imply that the discriminative and the non discriminative subspaces are orthogonal, which suggests in practice that all the relevant clustering information remains in the latent subspace. This model is referred to by $\text{DLM}_{[\Sigma_k \beta_k]}$ in [4] and a graphical summary is given in Figure 2.1.

2.2. A family of parsimonious model

Parsimonious models can be obtained by constraining the parameters Σ_k or β_k to be common between and within the groups. For instance, the covariance matrices $\Sigma_1, \ldots, \Sigma_K$ in the latent space can be assumed to be common across the groups and this submodel is referred to by $\text{DLM}_{[\Sigma \beta_k]}$. Similarly, in each group, Σ_k can be assumed to be diagonal, *i.e.* $\Sigma_k = \text{diag}(\alpha_{k1}, \ldots, \alpha_{kd})$. This submodel is referred to by $\text{DLM}_{[\alpha_{kj} \beta_k]}$. A constraint can also be applied in the parameter β_k by assuming it to be common to all classes ($\forall k$, $\beta_k = \beta$). This assumption can be viewed as modeling the non discriminative information with a unique parameter which seems natural for data obtained in a common acquisition process. A list of the 12 different DLM models is given by Table 1 and detailed descriptions can be found in [4]. Such a family yields very parsimonious models and allows, in the same time, to fit into various situations. In particular, the complexity of the $\text{DLM}_{[\Sigma_k \beta_k]}$ model mainly depends on the number of clusters K since the dimensionality of the discriminative subspace is such that $d \leq K - 1$. Notice that the complexity of the $\text{DLM}_{[\Sigma_k \beta_k]}$ grows linearly with p contrary to the traditional Gaussian models in which the complexity increases with p^2. As an illustration, if we consider the case where $p = 100$, $K = 4$ and $d = 3$, then the number of parameters to estimate for the $\text{DLM}_{[\Sigma_k \beta_k]}$ is 337 which

Model	Nb. of parameters	$K = 4$ and $p = 100$
$\text{DLM}_{[\Sigma_k\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K^2(K-1)/2+K$	337
$\text{DLM}_{[\Sigma_k\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K^2(K-1)/2+1$	334
$\text{DLM}_{[\Sigma\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K(K-1)/2+K$	319
$\text{DLM}_{[\Sigma\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K(K-1)/2+1$	316
$\text{DLM}_{[\alpha_{kj}\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K^2$	325
$\text{DLM}_{[\alpha_{kj}\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K(K-1)+1$	322
$\text{DLM}_{[\alpha_k\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+2K$	317
$\text{DLM}_{[\alpha_k\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K+1$	314
$\text{DLM}_{[\alpha_j\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+(K-1)+K$	316
$\text{DLM}_{[\alpha_j\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+(K-1)+1$	313
$\text{DLM}_{[\alpha\beta_k]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+K+1$	314
$\text{DLM}_{[\alpha\beta]}$	$(K-1)+K(K-1)+(K-1)(p-K/2)+2$	311
Full-GMM	$(K-1)+Kp+Kp(p+1)/2$	20603
Com-GMM	$(K-1)+Kp+p(p+1)/2$	5453
Mixt-PPCA	$(K-1)+Kp+K(d(p-(d+1)/2)+d+1)+1$	1198 ($d=3$)
Diag-GMM	$(K-1)+Kp+Kp$	803
Sphe-GMM	$(K-1)+Kp+K$	407

Table 1: Number of free parameters to estimate when $d = K-1$ for the DLM models and some classical models (see text for details).

is drastically less than in the case of the Full-GMM (20603 parameters to estimate).

2.3. The Fisher-EM algorithm

An estimation procedure, called the Fisher-EM algorithm, is also proposed in [4] in order to estimate both the discriminative space and the parameters of the mixture model. This algorithm is based on the EM algorithm from which an additional step is introduced, between the E and the M-step. This additional step, named F-step, aims to compute the projection matrix U whose columns span the discriminative latent space. The Fisher-EM algorithm has therefore the following form, at iteration q:

The E-step. This step computes the posterior probabilities $t_{ik}^{(q)}$ that the observations belong to the K groups using the following update formula:

$$t_{ik}^{(q)} = \hat{\pi}_k^{(q-1)}\phi(y_i, \hat{\theta}_k^{(q-1)}) / \sum_{\ell=1}^{K} \hat{\pi}_\ell^{(q-1)}\phi(y_i, \hat{\theta}_\ell^{(q-1)}), \tag{2.5}$$

with $\hat{\theta}_k = \{\hat{\mu}_k, \hat{\Sigma}_k, \hat{\beta}_k, \hat{U}\}$.

The F-step. This step estimates, conditionally to the posterior probabilities, the orientation matrix $U^{(q)}$ of the discriminative latent space by maximizing

the Fisher's criterion [9, 11] under orthonormality constraints:

$$\begin{aligned} \hat{U}^{(q)} &= \max_{U} \quad \text{trace}\left((U^t S U)^{-1} U^t S_B^{(q)} U\right), \\ &\quad \text{w.r.t.} \quad U^t U = \mathbf{I}_d, \end{aligned} \tag{2.6}$$

where S stands for the covariance matrix and $S_B^{(q)}$, defined as follows:

$$S_B^{(q)} = \frac{1}{n} \sum_{k=1}^{K} n_k^{(q)} (m_k^{(q)} - \bar{y})(m_k^{(q)} - \bar{y})^t, \tag{2.7}$$

denotes the soft between covariance matrix with $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}$, $m_k^{(q)} = 1/n_k^{(q)} \sum_{i=1}^{n} t_{ik}^{(q)} y_i$ and $\bar{y} = 1/n \sum_{i=1}^{n} y_i$. This optimization problem is solved in [4] using the concept of orthonormal discriminant vector developed by [10] through a Gram-Schmidt procedure. Such a process enables to fit a discriminative and low-dimensional subspace conditionally to the current soft partition of the data while providing orthonormal discriminative axes. In addition, according to the rank of the matrix $S_B^{(q)}$, the dimensionality of the discriminative space d is strictly bounded by the number of clusters K.

The M-step. This third step estimates the parameters of the mixture model in the latent subspace by maximizing the conditional expectation of the complete log-likelihood:

$$\begin{aligned} Q(\theta) = -\frac{1}{2} \sum_{k=1}^{K} n_k^{(q)} \Big[& \text{-}2 \log(\pi_k) + \text{tr}(\Sigma_k^{-1} \hat{U}^{(q)t} C_k^{(q)} \hat{U}^{(q)}) + \log(|\Sigma_k|) \\ & + (p\text{-}d) \log(\beta_k) + \frac{\text{trace}(C_k^{(q)}) \text{-} \sum_{j=1}^{d} \hat{u}_j^{(q)t} C_k^{(q)} \hat{u}_j^{(q)}}{\beta_k} + p \log(2\pi) \Big]. \end{aligned} \tag{2.8}$$

where $C_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} (y_i - \hat{m}_k^{(q-1)})(y_i - \hat{m}_k^{(q-1)})^t$ is the empirical covariance matrix of the kth group and $\hat{u}_j^{(q)}$ is the jth column vector of $\hat{U}^{(q)}$, $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$. Hence, maximizing Q conditionally to $\hat{U}^{(q)}$ leads to the following update formula for the mixture parameters of the model $\text{DLM}_{[\Sigma_k \beta_k]}$:

$$\hat{\pi}_k^{(q)} = \frac{n_k^{(q)}}{n}, \tag{2.9}$$

$$\hat{\mu}_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \hat{U}^{(q)t} y_i, \tag{2.10}$$

$$\hat{\Sigma}_k^{(q)} = \hat{U}^{(q)t} C_k \hat{U}^{(q)}, \tag{2.11}$$

$$\hat{\beta}_k^{(q)} = \frac{\mathrm{tr}(C_k) - \sum_{j=1}^{d} \hat{u}_j^{(q)t} C_k \hat{u}_j^{(q)}}{p-d}. \tag{2.12}$$

The Fisher-EM procedure iteratively updates the parameters until a stopping criterion is satisfied (see next paragraph). Finally, since the latent subspace has a low dimension and is also common to all groups, the clustered data can be easily visualized by projecting them into the estimated latent subspace.

2.4. Computational aspects

In all iterative procedures, both the initialization procedure and the stopping criterion have a significant effect on the algorithm performance. Regarding the initialization, several strategies have been proposed in the literature for initializing the EM algorithm. A popular practice [3], called mini-EM, executes the EM algorithm several times from a random initialization and only keeps the set of parameters associated with the highest likelihood. The use of k-means or of a random partition are also standard approaches for initializing the algorithm. In [4], it also suggested to initialize the Fisher-EM algorithm with the partition provided by the EM algorithm. On the other side, a classical stopping criterion is to stop the algorithm when the difference between two consecutive likelihood values is smaller than a positive value ε provided by the user. This stopping criterion will be used in the experiments of Section 4 and will be compared to an alternative proposed in this work. However, the stop of the algorithm with such a stopping criterion does not guarantee that it has reached a maximum of the likelihood.

3. Theoretical considerations on the convergence

The convergence of the Fisher-EM algorithm is first considered here from the theoretical point of view. Two cases are considered: the isotropic case (model $\mathrm{DLM}_{[\alpha\beta]}$) and the general case.

3.1. Isotropic case: model $DLM_{[\alpha\beta]}$

We first consider the model $\mathrm{DLM}_{[\alpha\beta]}$ which assumes a common and spherical covariance matrix for each class both in the latent subspace ($\forall k \in \{1,\ldots,K\}$, $\Sigma_k = \alpha \mathbf{I}_d$) and in its orthogonal complement ($\forall k \in \{1,\ldots,K\}$, $\beta_k = \beta$). Then, in this case, the following result holds.

Theorem 1. *In the case of the model DLM$_{[\alpha\beta]}$, the Fisher-EM algorithm is an EM algorithm and its convergence toward a local maximum of the likelihood is therefore guaranteed.*

Proof. In order to prove that the Fisher-EM algorithm is an EM algorithm in the case of the model DLM$_{[\alpha\beta]}$, it is necessary and sufficient to show that the maximization of the constrained Fisher's criterion (2.6) (involved in the F step) is equivalent to the maximization of the conditional expectation of the complete log-likelihood $Q(\theta)$ at iteration q.

On the one hand and by assuming that the empirical covariance matrix of the whole dataset is equal to I_p, the optimization problem (2.6) considered in the F step at iteration q can be rewritten, without loss of generality, as follows:

$$\begin{cases} \min\limits_{U} & \text{trace}\left(U^t S_W^{(q)} U\right), \\ \text{wrt} & U^t U = \mathbf{I}_d, \end{cases}$$

since $\text{cov(Y)} = S_B^{(q)} + S_W^{(q)}$ where $S_W = \frac{1}{n}\sum_{k=1}^{K} n_k^{(q)} C_k^{(q)}$ is the soft within covariance matrix, $C_k^{(q)} = \frac{1}{n_k^{(q)}}\sum_{i=1}^{n} t_{ik}^{(q)}(y_i - \hat{m}_k^{(q-1)})(y_i - \hat{m}_k^{(q-1)})^t$ is the empirical covariance of the kth group and $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$. In order to ease the reading, the index q of the current iteration is omitted in the remainder of the proof.

On the other hand, let us consider the quantity $-2Q(\theta)$ which has the following form in the case of the DLM models:

$$\begin{aligned} -2Q(\theta) &= -2\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}\log\left(\pi_k \phi\left(y_i;\theta_k\right)\right) \\ &= \sum_{k=1}^{K}\left[\sum_{i=1}^{n} t_{ik}[-2\log(\pi_k) + p\log\left(2\pi\right) + \log|S_k| + (y_i - m_k)^t S_k^{-1}(y_i - m_k)]\right] \\ &= \sum_{k=1}^{K}\left[\sum_{i=1}^{n} t_{ik}[\log|S_k| + (y_i - m_k)^t S_k^{-1}(y_i - m_k)]\right] + \gamma_1, \end{aligned}$$

where $\gamma_1 = \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}[-2\log(\pi_k) + p\log\left(2\pi\right)]$ is a constant term while maximizing with respect to U.

Let us now consider the case of the model DLM$_{[\alpha\beta]}$ which implies that $S_k = S = W\Delta W^t$, $\forall k \in \{1,\ldots,K\}$, and that the matrix Δ has the following form:

$$\Delta = \begin{bmatrix} \alpha\mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{I}_{p-d} \end{bmatrix}. \tag{3.1}$$

Given these assumptions, the quantity $\gamma_2 = \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}\log|S_k|$, which is also equal to $\sum_{k=1}^{K} n_k \log|S|$, is as well independent of U and then becomes a constant with respect to U.

Moreover, denoting by A the quantity $\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}(y_i - m_k)^t S^{-1}(y_i - m_k)$, we can state that:

$$\begin{aligned} A &= \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}(y_i - m_k)^t S^{-1}(y_i - m_k) \\ &= \text{trace}\left(S^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}(y_i - m_k)(y_i - m_k)^t\right) \\ &= n\text{trace}\left(S^{-1}S_W\right). \end{aligned}$$

Besides, since $S^{-1} = W\Delta^{-1}W^t$ where W satisfies $WW^t = W^tW = \mathbf{I}_p$, the quantity A can be rewritten as:

$$\begin{aligned} A &= n\text{trace}\left(\left(W^t\Delta W\right)^{-1}S_W\right) \\ &= n\text{trace}\left(\Delta^{-1}W^tS_WW\right). \end{aligned}$$

Let us finally introduce the matrices $\tilde{W} = [U, 0_{p-d}]$ and $\bar{W} = [0_d, V]$ such as $W = \tilde{W} + \bar{W}$, where V is an orthogonal complement of U. In this case, the relation $W^tS_WW = \tilde{W}^tS_W\tilde{W}^t + \bar{W}^tS_W\bar{W}$ can be easily stated since $\tilde{W}^tS_W\bar{W}$ and $\bar{W}^tS_W\tilde{W}$ are both null matrices. Therefore, according to the diagonal form of the matrix Δ (see equation (3.1)), then the quantity A becomes:

$$\begin{aligned} A &= n\text{trace}\left(\Delta^{-1}\left(\tilde{W}^tS_W\tilde{W}^t + \bar{W}^tS_W\bar{W}\right)\right) \\ &= n\left(\text{trace}\left(\frac{1}{\alpha}U^tS_WU\right) + \text{trace}\left(\frac{1}{\beta}V^tS_WV\right)\right) \\ &= \frac{n}{\alpha}\text{trace}\left(U^tS_WU\right) + \gamma_3, \end{aligned}$$

where $\gamma_3 = n\text{trace}\left(\frac{1}{\beta}V^tS_WV\right)$ is independent of U. Thus, the conditional expectation of the complete log-likelihood $Q(\theta)$ can be rewritten as:

$$-2Q(\theta) = \frac{n}{\alpha}\text{trace}\left(U^tS_WU\right) + \gamma,$$

where $\gamma = \gamma_1 + \gamma_2 + \gamma_3$.

Consequently, since minimizing the quantity $\text{trace}(U^tS_WU)$ with respect to U is equivalent to maximizing $Q(\theta)$, the F step of the Fisher-EM algorithm maximizes $Q(\theta)$ with respect to U in the case of the model $\text{DLM}_{[\alpha\beta]}$. This allows to conclude that the Fisher-EM algorithm, in the case of the model $\text{DLM}_{[\alpha\beta]}$, is a traditional EM algorithm and its convergence toward a local maximum of the likelihood is therefore guaranteed [17]. □

3.2. General case: all DLM models

We now consider the general case (all other models of the DLM family) and, in this case, the following result holds.

Theorem 2. *If, at each iteration q, the quantity*

$$\delta^{(q)} = \sum_{k=1}^{K} \text{trace}\left[n_k^{(q)} \left(\hat{\Sigma}_k^{(q-1)^{-1}} - \frac{1}{\hat{\beta}_k^{(q-1)}} \mathbf{I}_d \right) \left(\hat{U}^{(q-1)t} C_k^{(q)} \hat{U}^{(q-1)} - \hat{U}^{(q)t} C_k^{(q)} \hat{U}^{(q)} \right) \right]$$

is positive, then the Fisher-EM algorithm is a generalized EM (GEM) algorithm and its convergence toward a local maximum of the likelihood is therefore guaranteed.

Proof. In order to prove that the Fisher-EM algorithm is a generalized EM algorithm [8], it is necessary to show that, at each iteration q, $Q(U^{(q+1)}, \theta^{(q+1)}) \geq Q(U^{(q)}, \theta^{(q)})$, where $\theta^{(q)}$ is the set of model parameters estimated at iteration q, $U^{(q)}$ is the orientation matrix of the latent subspace and $Q(\theta)$ is the conditional expectation of the complete log-likelihood.

Let $\hat{U}^{(q)}$ and $\hat{\theta}^{(q)} = \left\{ \hat{\mu}^{(q)}, \hat{\mathbf{\Sigma}}^{(q)}, \hat{\beta}^{(q)}, \hat{\pi}^{(q)} \right\}$ be the model parameters estimated at iteration q and let $t_{ik}^{(q+1)}$, $i = 1, ..., n$ and $k = 1, ..., K$, be the posterior probabilities computed in the E step at iteration $q + 1$.

On the one hand, let us consider the quantity:

$$\delta^{(q+1)} = Q(\hat{U}^{(q+1)}, \hat{\theta}^{(q)}) - Q(\hat{U}^{(q)}, \hat{\theta}^{(q)}).$$

In the case of the DLM models, we recall that $Q(U, \hat{\theta}^{(q)})$ has the following form:

$$\begin{aligned} Q(U, \hat{\theta}^{(q)}) = -\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}^{(q+1)} \Big[-2 \log(\hat{\pi}_k^{(q)}) + \text{trace}\left(\left(\hat{\Sigma}_k^{(q)} \right)^{-1} U^t C_k^{(q+1)} U \right) + \log \left| \hat{\Sigma}_k^{(q)} \right| \\ + (p-d) \log(\hat{\beta}_k^{(q)}) + \frac{1}{\hat{\beta}_k^{(q)}} \text{trace}(C_k^{(q+1)} - U^t C_k^{(q+1)} U) + p \log(2\pi) \Big]. \end{aligned}$$

where $C_k^{(q+1)}$ is the empirical covariance matrix of the kth group computed at iteration $q + 1$ (conditionally to the posterior probabilities $t_{ik}^{(q+1)}$). By subtracting term by term, we end up with:

$$\delta^{(q+1)} = \frac{1}{2} \left[\sum_{k=1}^{K} \text{trace}\left(B_k^{(q)} \left(A_k^{(q)} - A_k^{(q+1)} \right) \right) \right],$$

where:

$$\begin{aligned} A_k^{(q)} &= \hat{U}^{(q)t} n_k^{(q+1)} C_k^{(q+1)} \hat{U}^{(q)} \\ A_k^{(q+1)} &= \hat{U}^{(q+1)t} n_k^{(q+1)} C_k^{(q+1)} \hat{U}^{(q+1)} \\ B_k^{(q)} &= \hat{\Sigma}_k^{(q)^{-1}} - \frac{1}{\hat{\beta}_k^{(q)}} \mathbf{I}_d. \end{aligned}$$

Although the criterion maximized in the F step guarantees that the quantity $\sum_{k=1}^{K} \text{trace}\left(A_k^{(q)} - A_k^{(q+1)}\right) \geq 0$ if $S = I_p$, we have however no guarantee that $\text{trace}\left(A_k^{(q)} - A_k^{(q+1)}\right) \geq 0$ for all $k = 1, ..., K$. It is therefore not possible to be sure that, at each iteration, $\delta^{(q+1)} \geq 0$ even though $B_k^{(q)}$ is a semi-definite positive matrix. In order to go further, let us therefore assume that the following condition is satisfied:

$$\text{H1}: \ \delta^{(q+1)} = \frac{1}{2}\left[\sum_{k=1}^{K} \text{trace}\left(B_k^{(q)}\left(A_k^{(q)} - A_k^{(q+1)}\right)\right)\right] \geq 0.$$

On the other hand, the EM algorithm theory [8] implies that the set of parameter estimates $\hat{\theta}^{(q+1)} = \left\{\hat{\mu}^{(q+1)}, \hat{\mathbf{\Sigma}}^{(q+1)}, \hat{\beta}^{(q+1)}, \hat{\pi}^{(q+1)}\right\}$ (estimated in the M step at iteration $q+1$) is such that $Q(\hat{U}^{(q+1)}, \hat{\theta}^{(q+1)}) \geq Q(\hat{U}^{(q+1)}, \theta)$ for any θ.

It is now straightforward to conclude since, in particular, $Q(\hat{U}^{(q+1)}, \hat{\theta}^{(q+1)}) \geq Q(\hat{U}^{(q+1)}, \hat{\theta}^{(q)})$ and $Q(\hat{U}^{(q+1)}, \hat{\theta}^{(q)}) \geq Q(\hat{U}^{(q)}, \hat{\theta}^{(q)})$ if Assumption H1 is verified. Consequently, conditionally to the fact that H1 holds, the Fisher-EM algorithm is a generalized EM algorithm and its convergence toward a local maximum of the likelihood is therefore guaranteed [17] in the general case. □

The convergence condition H1 seems however not to be a strong conditions since, as we said before, the criterion maximized in the F step implies that $\sum_{k=1}^{K} \text{trace}\left(A_k^{(q)} - A_k^{(q+1)}\right) \geq 0$ at each iteration q and since $B_k^{(q)}$ is a semi-definite positive matrix. We therefore believe that H1 is frequently satisfied in practice. In addition, it is easy to monitor the quantity $\delta^{(q)}$ along the iterations to verify if H1 is satisfied for the clustering task at hand. Such a verification is made on a real-world dataset in the following section.

4. Practical considerations on the convergence

We now focus on the practical aspects of the Fisher-EM convergence. We first present an experimental validation of the convergence criterion intro-

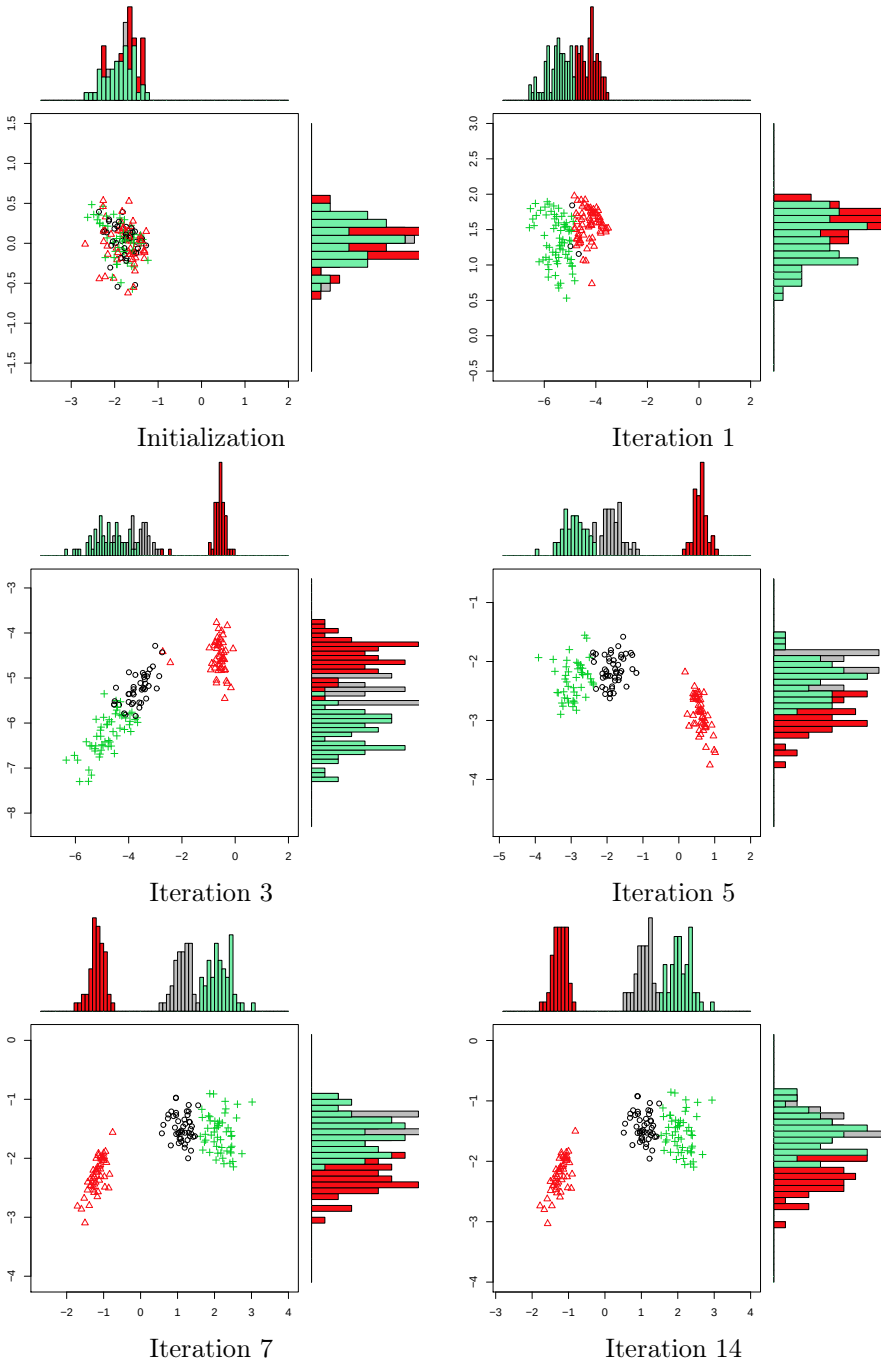

Figure 4.1: Projection of clustered Iris data into the estimated latent discriminative subspace at some iterations of the Fisher-EM algorithm.

duced in Theorem 2. The use of the Fisher's criterion as stopping criterion is then investigated and Fisher-EM is finally compared to the EM and CEM algorithms.

4.1. Experimental validation

The Iris dataset is used here as an introductory example because of the link with Fisher's work [9] but also for its popularity in the clustering and classification communities. This dataset, collected by E. Anderson [1] in the Gaspé peninsula (Canada), is made of three groups corresponding to different species of iris (*setosa*, *versicolor* and *virginica*) among which the groups *versicolor* and *virginica* are difficult to discriminate (they are at least not linearly separable). The dataset consists of 50 samples from each of three species and four features were measured from each sample. The four measurements are the length and the width of the sepal and the petal.

For this experiment, we used the Fisher-EM algorithm with the model $\text{DLM}_{[\alpha_{kj}\beta_k]}$ to cluster the 150 observations into three groups. The labels have been of course used only for the evaluation of the clustering performance. The algorithm was initialized with a random partition drown from a multinomial distribution with equal prior probabilities. Figure 4.1 shows the projection of clustered data into the estimated latent discriminative subspace at some iterations of the Fisher-EM algorithm. The current partition of the data is indicated at each iteration by the colors. Group-specific histograms provide as well some information on the projected distributions of the groups on each axis. It can be observed that the estimated latent space of the last iteration discriminates almost perfectly the three different groups. For this experiment, the clustering accuracy has reached 98%.

Figure 4.2 presents the evolution of the log-likelihood and of the convergence criterion $\delta^{(q)}$ (*cf.* Theorem 2) according to the iterations of the Fisher-EM algorithm on the Iris dataset. As expected, one can observe that the convergence criterion $\delta^{(q)}$ is large at the beginning and decreases toward 0 when the likelihood reaches a stationary value. On this example, it is therefore guaranteed that the Fisher-EM algorithm reached a local optimum of the likelihood.

4.2. The Fisher's criterion as stopping criterion

The Fisher-EM algorithm iteratively maximizes two quantities, the likelihood and the constrained Fisher's criterion (2.6), and, as shown by the

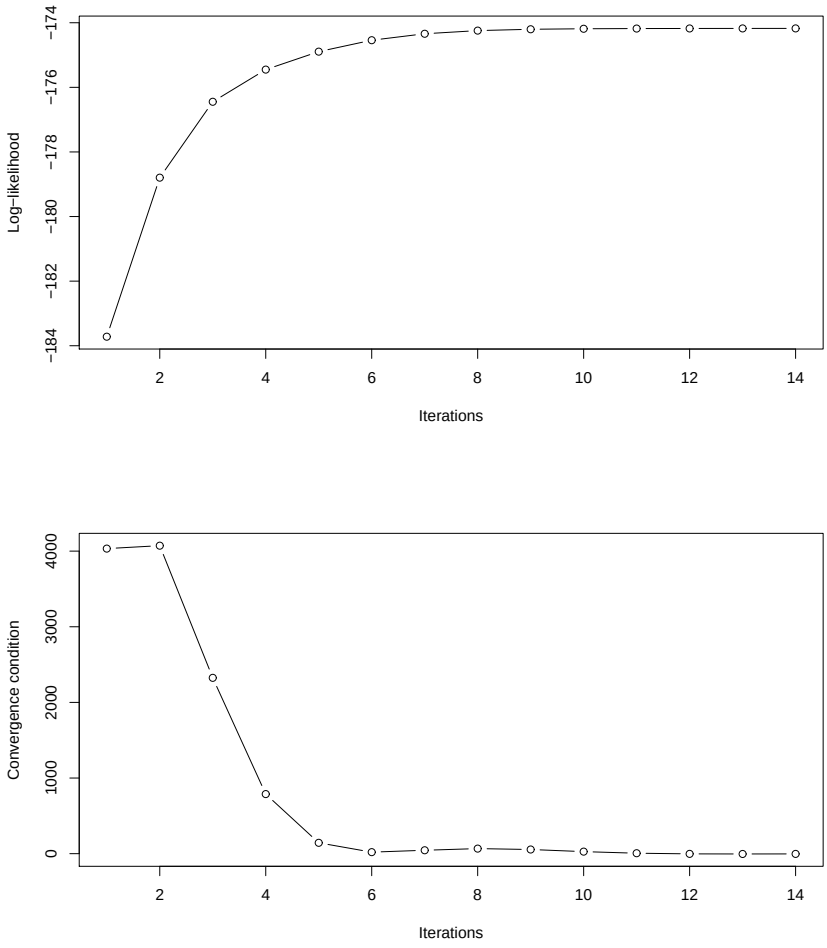

Figure 4.2: Evolution of the log-likelihood (top) and of the convergence criterion $\delta^{(q)}$ (bottom) according to the iterations of the Fisher-EM algorithm on the Iris dataset.

theoretical study of Section 3, both quantities have a strong relationship. Furthermore, the Fisher-EM algorithm classically stops when the difference between two consecutive likelihood values is smaller than a positive value ε provided by the user. It is therefore questionable if the Fisher's criterion can be used as stopping criterion instead of the likelihood. This experiment aims to answer this question.

To that end, we simulated a dataset made of 300 observations coming from 3 groups (with equal prior probabilities) in a 25-dimensional space according to the model $\text{DLM}_{[\alpha_k\beta]}$. The dimension of the latent space was $d = 2$ and the transformation matrix $W = [U, V]$ was randomly simulated such as $W^tW = WW^t = I_p$. The group means and the noise variance were set up such that the clustering problem was difficult. The used parameters were $\mu_1 = (0, 0)$, $\mu_2 = (0, 3)$, $\mu_3 = (-3, 0)$, $\alpha_1 = 0.25$, $\alpha_2 = 1$, $\alpha_3 = 0.5$ and $\beta = 1$.

Figure 4.3 shows the evolution of the log-likelihood and of the Fisher criterion according to the Fisher-EM iterations for the clustering of a simulated dataset. It clearly appears that the likelihood reaches a stationary value faster than the Fisher's criterion. According to these behaviors and if a stopping criterion with $\varepsilon = 1e-3$ is applied on both the standardized log-likelihood and Fisher's criteria, the algorithm stops after 9 iterations when considering the likelihood and after 38 iterations when considering the Fisher's criterion. The difference between both criteria can be explained by the fact that the likelihood is mostly associated with the fitting quality whereas the Fisher's criterion is more related to the group separation and consequently to the clustering accuracy. On this example, the parameter estimation turns out to be satisfying long before the group separation.

This difference can be quantified by computing the clustering accuracy associated with the clustering results obtained with both criteria. Figure 4.4 shows the evolution of the clustering accuracy according to the Fisher-EM iterations for the simulated dataset. The Fisher-EM algorithm stops at the red solid line, after 9 iterations, if the log-likelihood is used and at the green dashed line, after 38 iterations, if the Fisher criterion is used (when $\varepsilon = 1e-3$). From this figure, the Fisher's criterion seems to be a more reliable stopping criterion than the likelihood when considering the clustering task.

In order to validate this observation, we computed both the average number of iterations and clustering accuracy for both the likelihood and the

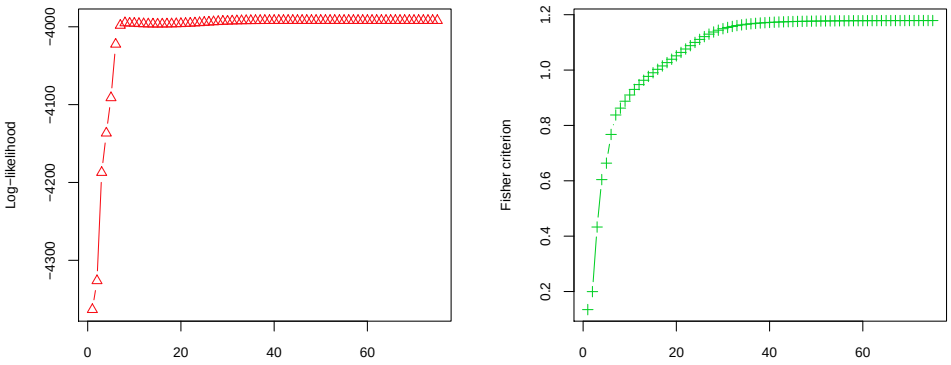

Figure 4.3: Evolution of the log-likelihood (left) and of the Fisher criterion (right) according to the Fisher-EM iterations for the clustering of a simulated dataset.

Fisher's criterion on 25 replications of the experiment. Figure 4.5 presents these results. It clearly appears that the use of the Fisher's criterion as stopping criterion for the Fisher-EM algorithm yields to a significant larger number of iterations but also to a significant higher clustering accuracy compared to the likelihood. To summarize, this experiment has shown that, when looking more for a high clustering accuracy than a good parameter estimation, it is preferable to consider the constrained Fisher's criterion (2.6) as stopping criterion for the Fisher-EM algorithm.

4.3. Fisher-EM vs. EM and CEM algorithms

We focus now, still from the practical point of view, on the convergence rate of the EM, classification EM (CEM) [6] and Fisher-EM algorithms. The convergence rate of the EM algorithm is known to be relatively slow. Dempster *et al.* [8] show that the rate of convergence of the EM algorithm is linear and that it depends on the proportion of information in the observed data. In order to fasten the convergence rate of the EM algorithm when the practitioner is mostly interested in the clustering performance, Celeux and Govaert [6] proposed the CEM algorithm which adds a classification step between the E and M step. The CEM algorithm is in particular known for converging faster than the EM algorithm.

This experiment aims to compare, in a simulation setup, the convergence rate of the Fisher-EM algorithm with the ones of the EM and CEM algorithms. To that end, we simulated a dataset made of 600 observations coming from 3 groups (with equal prior probabilities) in a 5-dimensional space according to the model $\text{DLM}_{[\alpha_k\beta]}$. The dimension of the latent space was $d = 2$

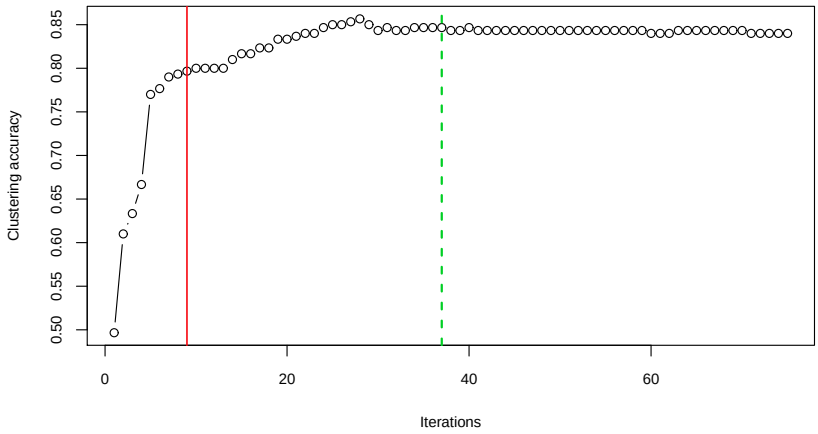

Figure 4.4: Clustering accuracy according to the Fisher-EM iterations for the simulated dataset. The Fisher-EM algorithm stops at the red solid line if the log-likelihood is used as stopping criterion and at the green dashed line if the Fisher criterion is used (both with $\varepsilon = 1e-3$).

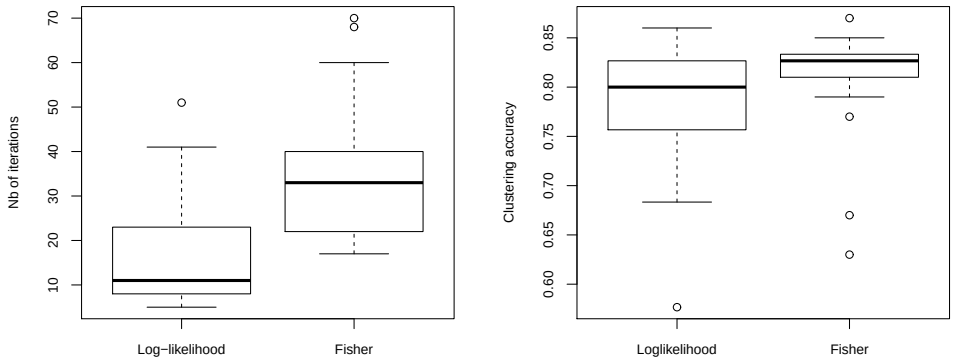

Figure 4.5: Average behaviors of the likelihood and the Fisher criterion as stopping criteria for the clustering of a simulated dataset (25 replications). The left panel shows the average number of iterations and the right panel shows the resulting average clustering accuracy.

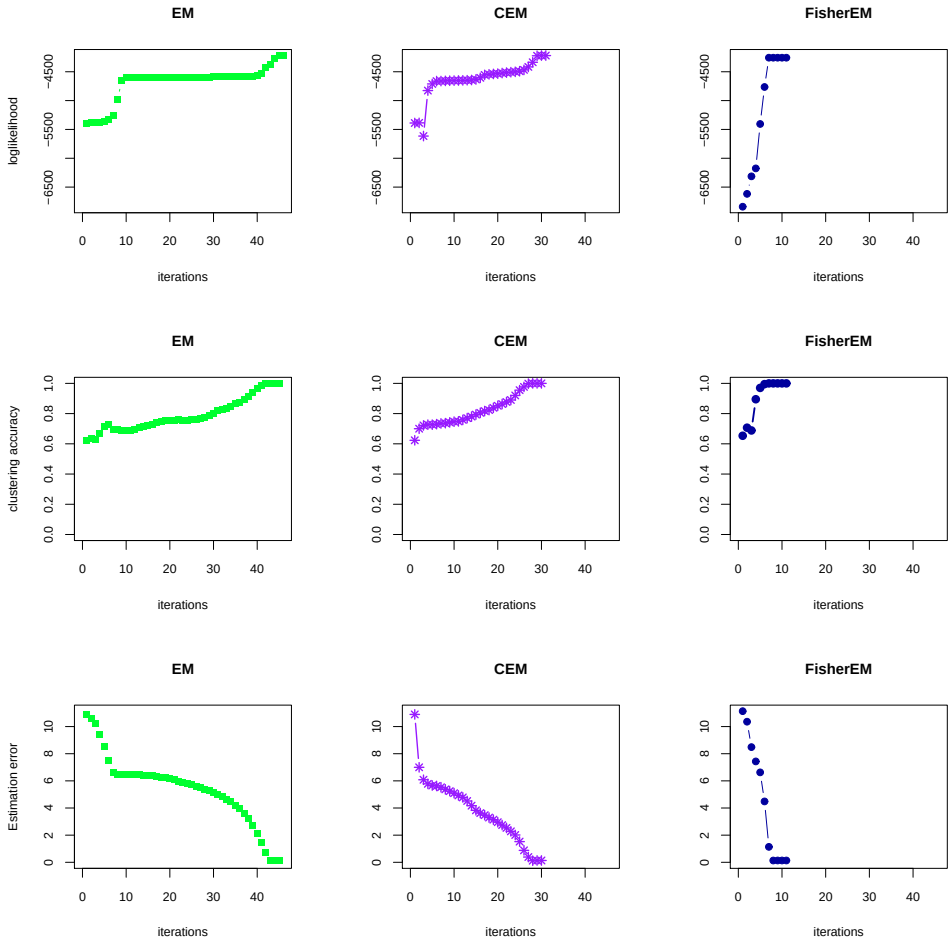

Figure 4.6: Evolution of the log-likelihood (top), the clustering accuracy (center) and the estimation error (bottom) according to the number of iterations for the EM, CEM and Fisher-EM algorithm.

and the transformation matrix $W = [U, V]$ was once again randomly simulated such as $W^tW = WW^t = I_p$. Conversely to the previous experiment, the group means and the noise variance were set up such that the clustering problem was easy. The used parameters were $\mu_1 = (0, 0)$, $\mu_2 = (0, 10)$, $\mu_3 = (-10, 0)$, $\alpha_1 = 0.25$, $\alpha_2 = 1$, $\alpha_3 = 0.5$ and $\beta = 0.5$. Hence, the EM and CEM algorithms are not disadvantaged and, given the low dimension of the data and the large number of observations, they should be able to correctly fit the data. For the three algorithms, the stopping criterion is based on the likelihood and the same $\varepsilon = 1e - 6$ is used in all cases.

Figure 4.6 and 4.9 present the evolution of the log-likelihood, the cluster-

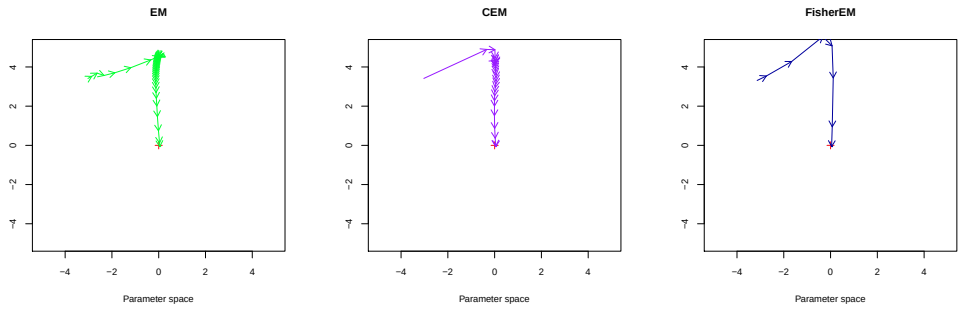

Figure 4.7: Estimation path in the parameter space (latent mean of group #1) for the EM (left), CEM (center) and Fisher-EM algorithm (left). Each arrow represents one iteration of the algorithm.

ing accuracy and the estimation error according to the number of iterations for the EM, CEM and Fisher-EM algorithm. The estimation error $e^{(q)}$ at iteration q was computed only on the group means according to the following formula: $e^{(q)} = \sum_{k=1}^{K} \|\mu_k - \hat{\mu}_k^{(q)}\|^2$. Firstly, as shown by the final likelihood values of the three algorithms, the simulated dataset seems quite easy to model. Indeed, the EM and CEM algorithms end up with likelihood values closed to the one of the Fisher-EM algorithm for which the model has been used to simulate the data. As expected, the CEM algorithm converges faster than the EM algorithm but provides similar results for the clustering and the parameter estimation. The Fisher-EM algorithm turns out to converge faster than both the EM and CEM algorithms without any deterioration of the clustering and estimation results.

Figure 4.7 shows the estimation path of each algorithm in the parameter space (mean of the 1st group). The actual value of the parameter is indicated by the red plus at the center of each panel. Each arrow represents one iteration of the algorithm. It also appears here that the Fisher-EM algorithm is more efficient than both the EM and CEM algorithms in finding the actual value of the parameter in the parameter space.

Finally, Figure 4.8 presents the average number of iterations, clustering accuracy and estimation error for the EM, CEM and Fisher-EM algorithms on 25 replications of the experiment. These results confirm that the Fisher-EM algorithm converges faster than both the EM and CEM algorithms while providing similar or better clustering and estimation performances.

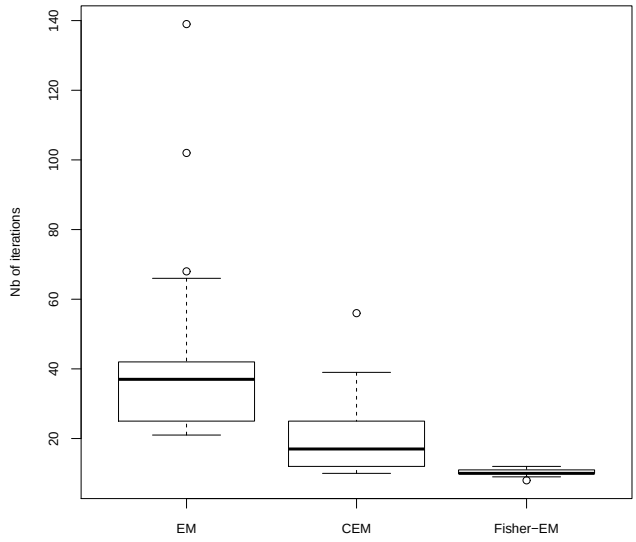

Figure 4.8: Average number of iterations for the EM, CEM and Fisher-EM algorithms (25 replications).

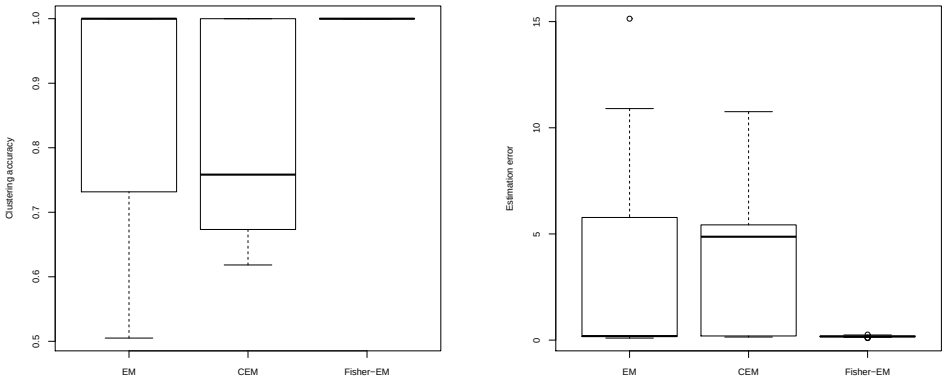

Figure 4.9: Average clustering accuracy (left) and estimation error (right) for the EM, CEM and Fisher-EM algorithms (25 replications). .

5. Conclusion

This article has focused on the convergence properties of the Fisher-EM algorithm, which has been recently proposed for the simultaneous visualization and clustering of high-dimensional data. The aim of this work was two folds. Firstly, the convergence of the Fisher-EM algorithm toward a local optimum of the likelihood has been proved in the isotropic case. The convergence has been proved as well in the general case under a weak condition which is easy to monitor in practice. Secondly, the convergence of the Fisher-EM algorithm has been studied from a practical point of view. Numerical experiments have in particular shown that the Fisher's criterion can be used as stopping criterion when considering mainly the clustering goal. It has been also shown that the Fisher-EM algorithm converges faster than both the EM and CEM algorithm.

Among the possible extensions of this work, it could be interesting to propose a unified estimation procedure for both the orientation matrix U and the other model parameters. This should be at least possible in the isotropic case since we showed that, in this case, the maximization of the Fisher's criterion is equivalent to the maximization of the likelihood. Another interesting extension would be to modify the F step such that the convergence criterion of Theorem 2 is always satisfied in the general case.

References

[1] E. Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

[2] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[3] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.

[4] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, to appear:1–24, 2011.

[5] C. Bouveyron, S. Girard, and C. Schmid. High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.

[6] G. Celeux and G. Govaert. A Classification EM Algorithm for Clustering and Two Stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

[7] W.C. Chang. On using principal component before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society, Series C*, 32(3):267–275, 1983.

[8] A. Dempster, N. Laird, and D. Robin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[9] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[10] D.H. Foley and J.W. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computers*, 24:281–289, 1975.

[11] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic. Press, San Diego, 1990.

[12] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, (41):379, 2003.

[13] P. McNicholas and B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

[14] P. McNicholas and B. Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

[15] A. Montanari and C. Viroli. Dimensionally reduced mixtures of regression models. *Electronic Proceedings of KNEMO, Knowledge Extraction and Modelling*, 2006.

[16] A. Montanari and C. Viroli. Heteroscedastic Factor Mixture Analysis. *Statistical Modeling: An International journal (forthcoming)*, (to appear), 2010.

[17] C. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

[18] R. Yoshida, T. Higuchi, and S. Imoto. A mixed factor model for dimension reduction and extraction of a group structure in gene expression data. *IEEE Computational Systems Bioinformatics Conference*, 8:161–172, 2004.